INACH

Bringing the Online In Line with Human Rights

licra

INACH Monitoring Report 2022

Compiled by
Adinde Schoorl and Maia Feijoo
Edited by
Tamás Berecz

2022

Table of Contents

1) Basic information about the monitoring exercise 1
2) Findings of the ME 1
3) Types of hate and intersectionality 4
4) IT platforms performances and observations from NGOs 4
5) Conclusion 5

1) Basic information about the monitoring exercise

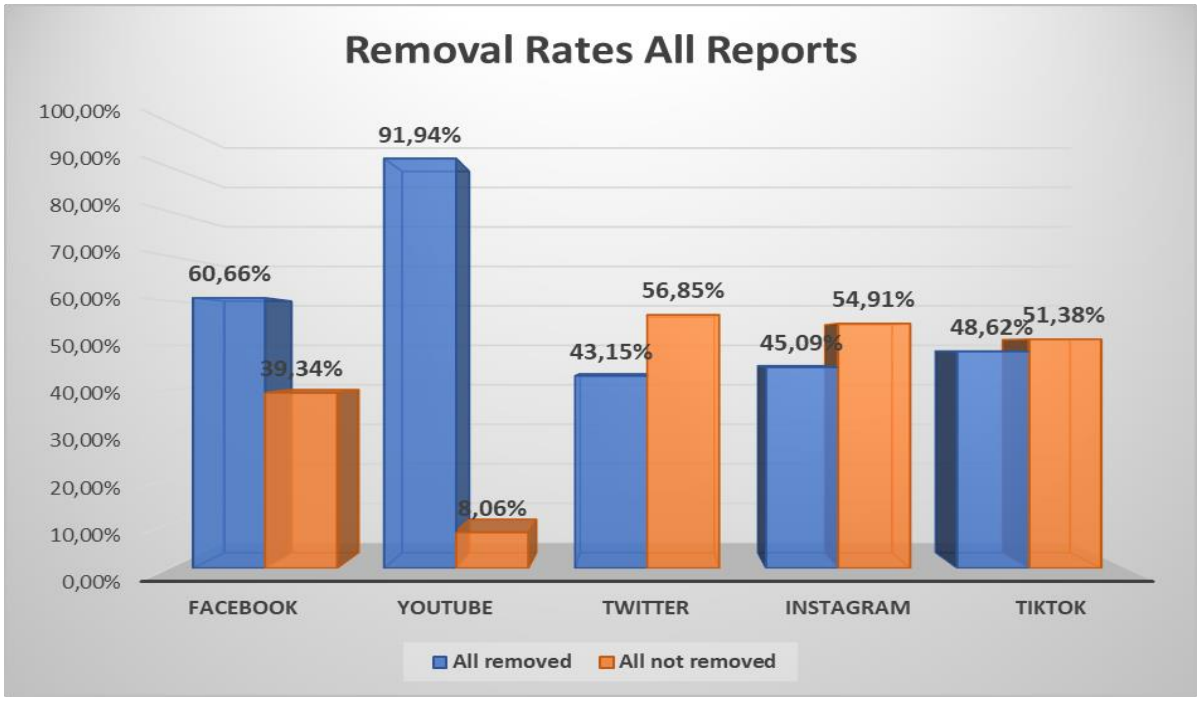
This seventh monitoring exercise was organized during six weeks, between the 28th of March and the 13th of May 2022, and coordinated by INACH and Licra. In total, nineteen NGOs took part in the exercise, supervised by the European Commission: Active Watch, DigiQ, Latvian Center for Human Rights, CESIE, Fundación Secretariado Gitano, Jugendschutz.net, Hatter Society, Human Rights House Zagreb, ILGA, Never Again Association, ROMEA, ZARA, LGL, Greek Helsinki Monitor, Estonian Human Rights Center, Integro Association, Institutet for juridic och internet, LICRA and INACH.

The objective of this exercise was to report illegal online content on the IT platforms signatories of the Code of Conduct on Online Hate and analyse how the social media moderate it. The NGOs focused on the types of hate speech they reported, if there were cases containing intersectionality, the removal and assessment rates from the platforms.

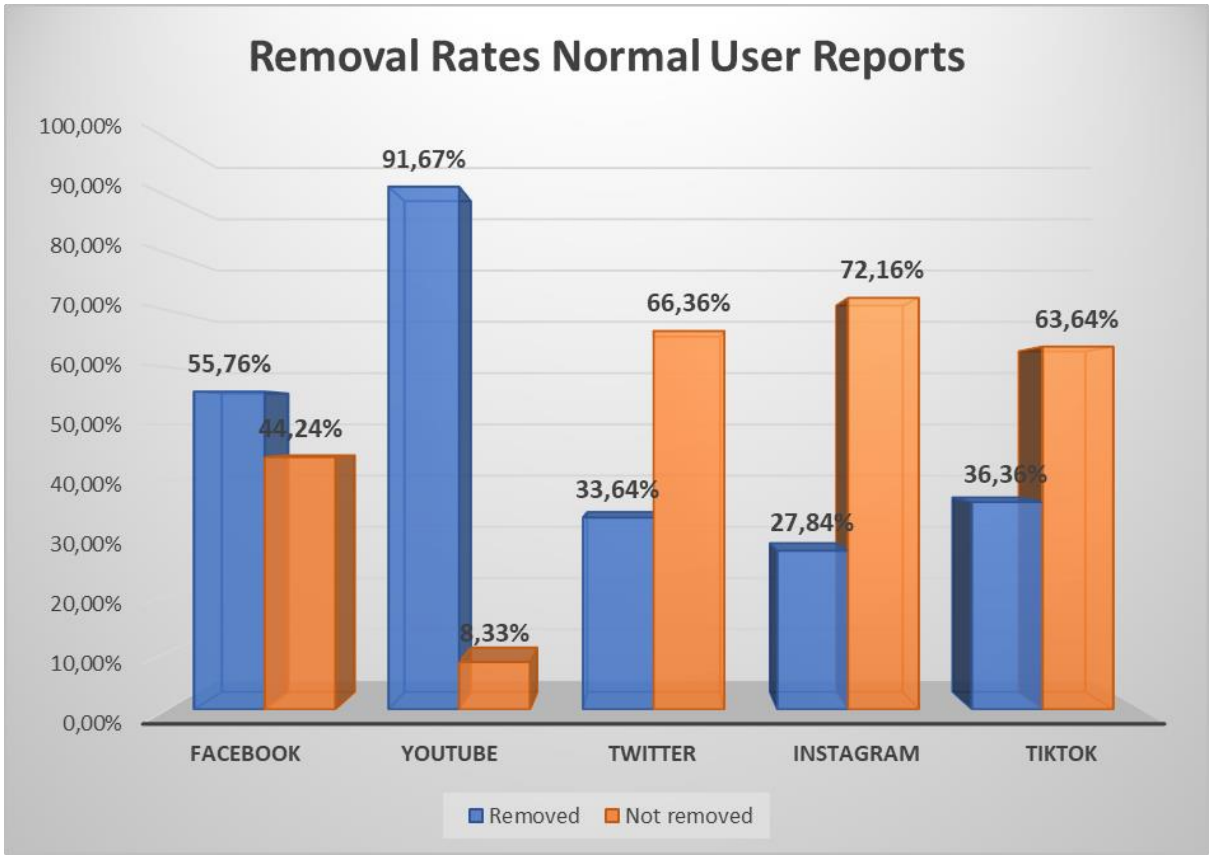
In total, the NGOs reported more than 1900 cases to the signatory platforms.

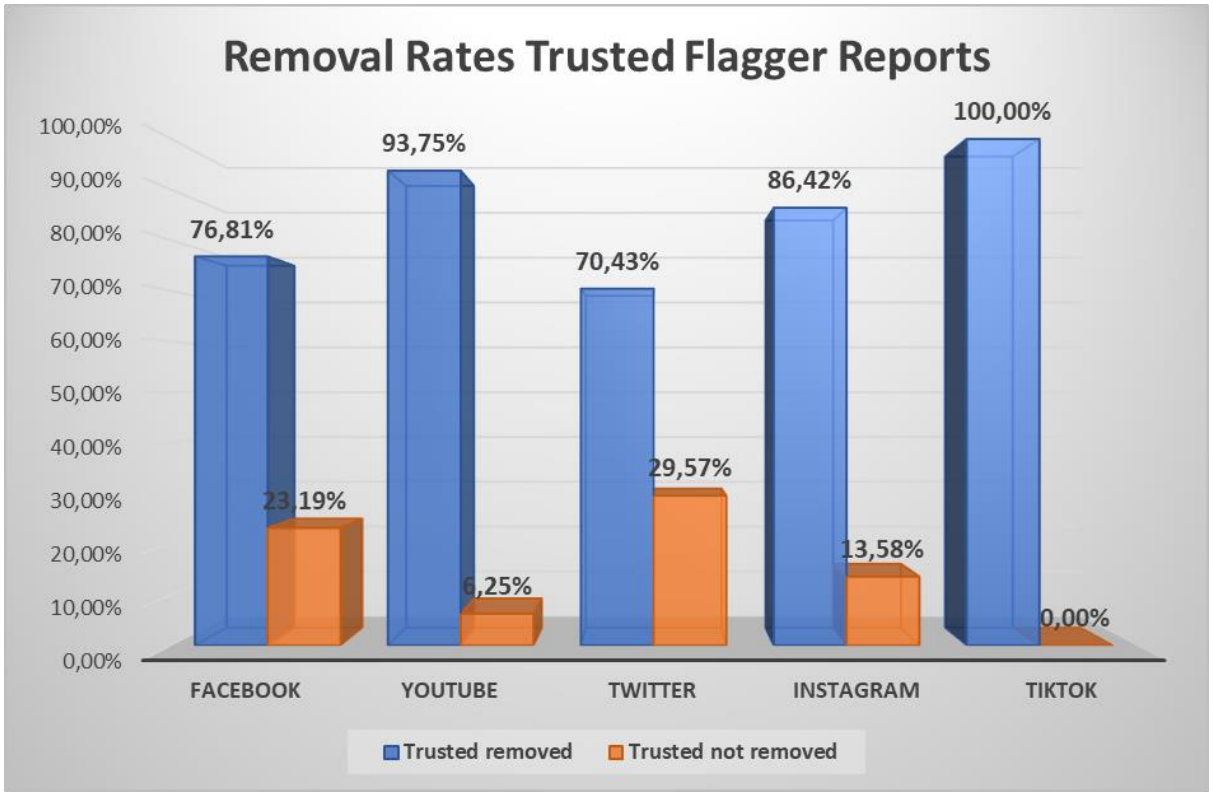
2) Findings of the ME

We observe that the results of the monitoring exercise have changed compared to last year. Overall, the removal rate decreased compared to last year. Almost all Social Media Companies have a lower removal rate compared to last year. The only outlier is YouTube; it has an impressive, improved removal rate of 91.94%, compared to a removal rate of 58.8 % last year.

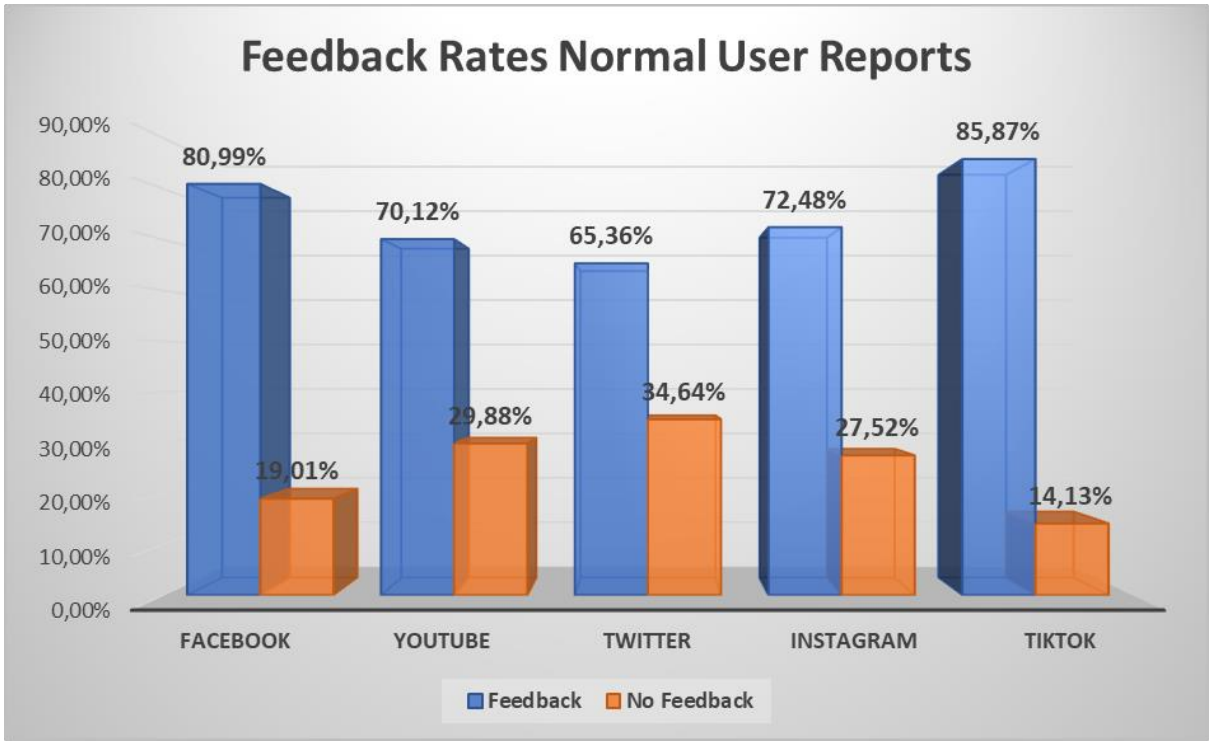


With this Monitoring Exercise we have also been able to generate data on the difference between removal rates as a Normal User and as a Trusted Flagger and the graphs below show interesting data. There is a clear difference between the removal rate as a Normal User and as a Trusted Flagger because all Social Media Companies have a higher removal rate when reporting is done by a Trusted Flagger.

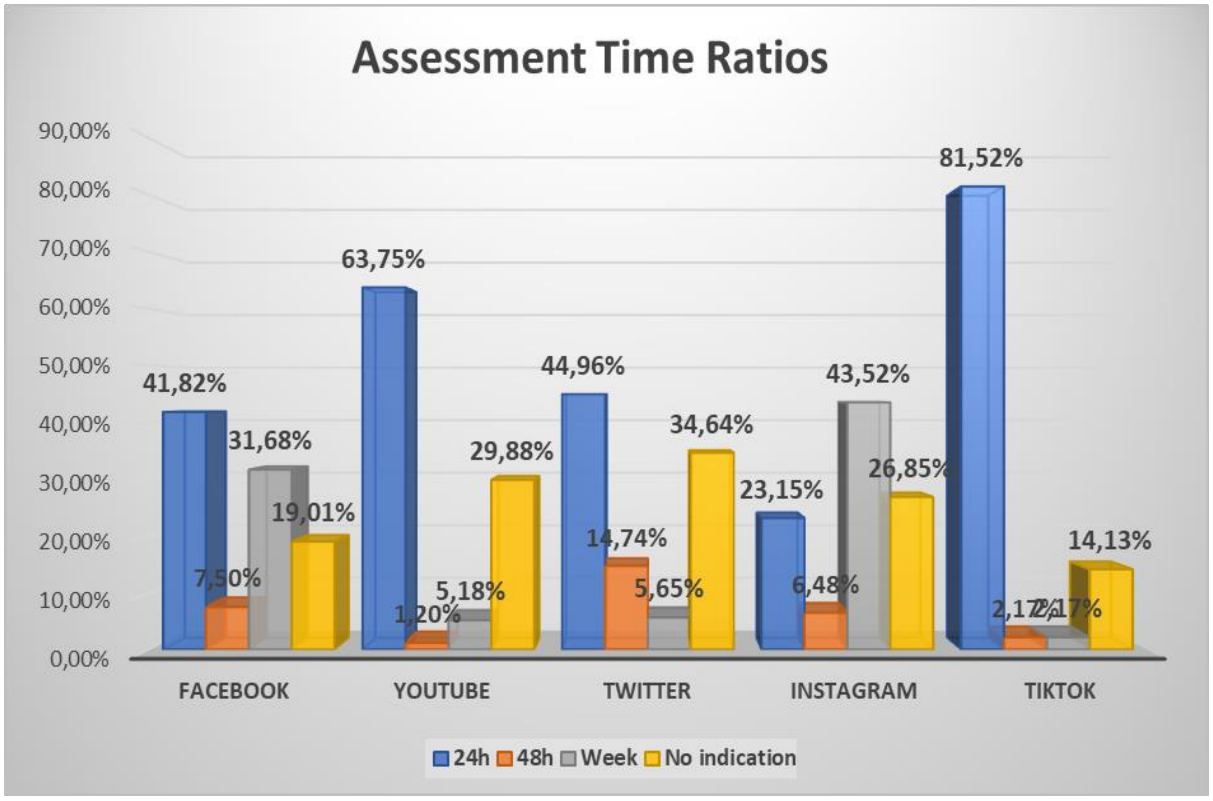




The feedback rate increased in the case of almost all the Social Media Companies, except for Facebook. The feedback rate for Facebook decreased from 86.9 % to 76.81 %. Especially YouTube had a dramatic increase in the feedback rate; from 7.3% in 2021 to 70.12% in 2022.

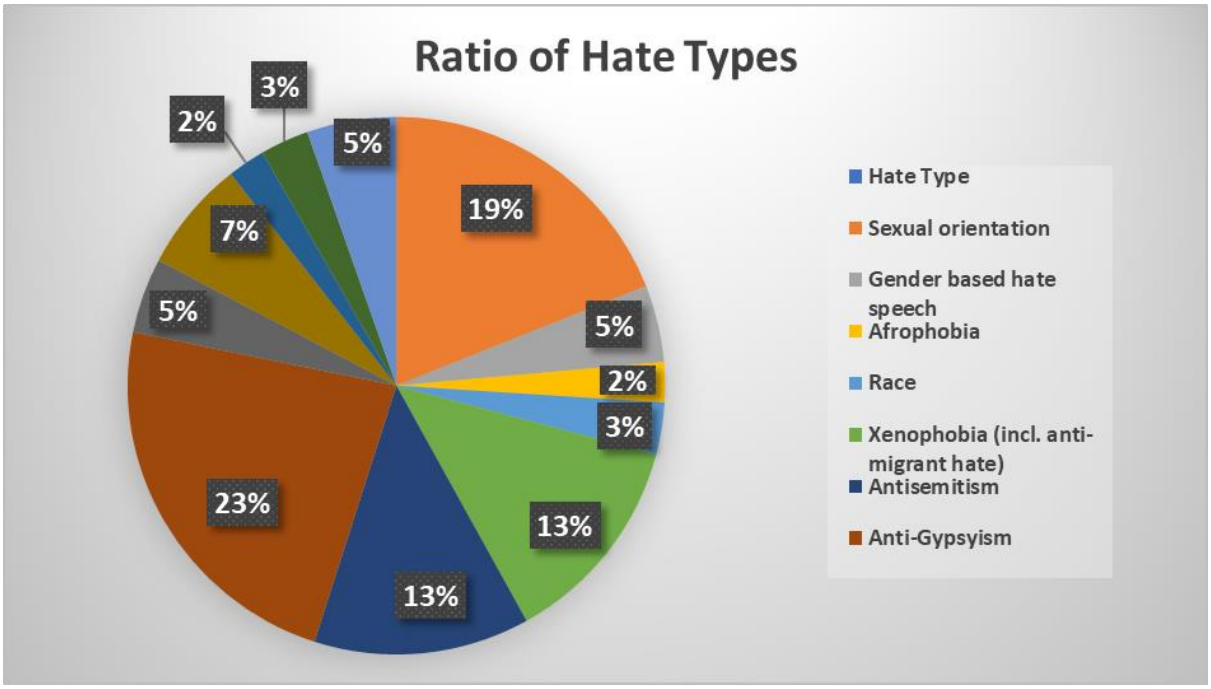


The majority of the responses of Social Media Companies have been within a 24 hours of assessment ratio. Especially TikTok has a high response rate within 24 hours; 81.52 %. However, it is disappointing that there is still a large percentage of reports that only receive a response within one week. And compared to last year all Social Media Companies have a lower percentage of assessment ratio within 24 hours.



3) Types of hate and intersectionality

The most prevalent types of hate collected during this exercise were: Antigypsyism, antisemitism, xenophobia (including anti-migrant) and anti-LGBT.
It is important to highlight the fact that some of the NGOs taking part in the ME are only focusing on Antigypsyism or anti-LGBT hate speech. We also noticed that the prevalence of antisemitic and xenophobic content can be arguably explained by the ongoing war in Ukraine, the conspiracy theory according to which Jews are perpetrating this conflict, and the refugee flows it is engendering.



The NGOS also found some cases of intersectionality, especially cases of Antigypsyism linked to anti-migrant and / or misogynist hate speech, cases of ethnic and national hatred, and cases of anti-migrant related to antisemitic hate speech. Intersectionality is a notion used in sociology and political reflexion, referring to the situation of a person suffering different forms of discrimination at the same time. According to the European Institute for Gender Equality, intersectionality is an "analytical tool for studying, understanding and responding to the ways in which sex and gender intersect with other personal characteristics / identities, and how these intersections contribute to unique experiences of discrimination".

4) IT platforms performances and observations from NGOs

In the overall perception of the results of the ME, we noticed that the removal rate is lower than the previous year, although the assessment from the platforms is higher. In general, the NGOs highlighted the fact that when they used their trusted flagger status to report a content, the performances of the IT platforms were better in the removal and the 24-hour assessment.

The company that most removed content is YouTube, with a total of 91,94%. Twitter has the lowest removal rate, with a total of 56,85%. It is also important to notice that the removal rate for normal users is still the highest with YouTube (91,67%), whereas Instagram is the lowest (72,16%). Regarding the removal rate with trusted flagger status, TikTok is the highest (100%) and Twitter is the lowest (29,57%).

TikTok has the highest 24-hour time of assessment, with a total of 81,52%, whereas in 34,64% of the cases, Twitter never sent any feedback. We highlight that in some cases, Twitter removed content without notifying the user, so we don't know if the report has been considered by this company or not.

We also noticed that Instagram has the lowest rate of 24-hour assessment (23,15%). At least three NGOS notified that, after a normal user report, Instagram sent them feedback explaining that the content would not be reviewed by their moderation system, due to a high number of reports. This means that the NGOS had to use their trusted flagger status because this company would not check normal user reports.

5) Conclusion

In conclusion to the 2022 main monitoring exercise, we have to – sadly – make the observation that almost all major platforms did worse than in the previous main exercise. They removed less cases and they provided timely assessment in less cases. The only outlier is YouTube; it has improved its removal rate of 91.94%, compared to a removal rate of 58.8 % last year, an impressive shift.

The only positive change we can observe is a major shift in providing feedback to reports. All platforms showed fundamental and impressive improvement. This might be due to their preparations to adhere to the DSA, which makes providing substantial feedback compulsory to these platforms.

Furthermore, there are still whopping discrepancies between the removal rates of normal user reports and trusted flagger reports. This issue has been highlighted by INACH and our members countless times since the start of the MEs. As we wrote in our previous report: "From the very first monitoring in 2016, NGOs observed that there are major differences in removal rates when a certain piece of content is reported via the normal user channel or via the trusted flagger channels that are only open to hate speech experts. At first glance this seems logical and probably even the platforms would argue that this is a good thing. They utilise this trusted flagger system to harness the expertise of NGOs and thus, they take the reports coming through these channels more seriously. However, quite often content that does not get removed after a normal user report, but will get removed after it gets reported via a trusted flagger channel. This is simply unacceptable. Content should be removed because it is illegal, not because it was reported by an expert instead of an everyday user. A piece of content is either illegal hate speech or not, it is either in breach of the community guidelines or not. Hence, there should be basically no or very little difference between removal of normal user and trusted flagger reports." Again, there was an outlier to this issue, and again it was YouTube. This platform only had a 2 percent difference between the removal rates of normal users and trusted flagger reports.

Overall, we can say that YouTube has gone through a fundamental shift in its approach to moderating illegal hate speech on their platform and they improved on all their numbers from the previous years.